Groupwork2

Kuan

2023-03-10

```
library(readr)
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(janitor)
library(areaplot)
library(dplyr)
library(skimr)
library(kableExtra)
library(gridExtra)
library(ggplot2)
library(stringr)
library(ISLR)
library(plotly)
library(MASS)
library(broom)
```

Data pre-processing

Remove missing value

In the raw dataset, there are some missing data about mean altitude and harvested, so before analysis data we remove missing values.

```
dataset13 <- read_csv("dataset13.csv")
dim(dataset13)
```

```
[1] 1145    8
```

```
#remove NA
newdataset<- na.omit(dataset13)
dim(newdataset)
```

```
[1] 935   8
```

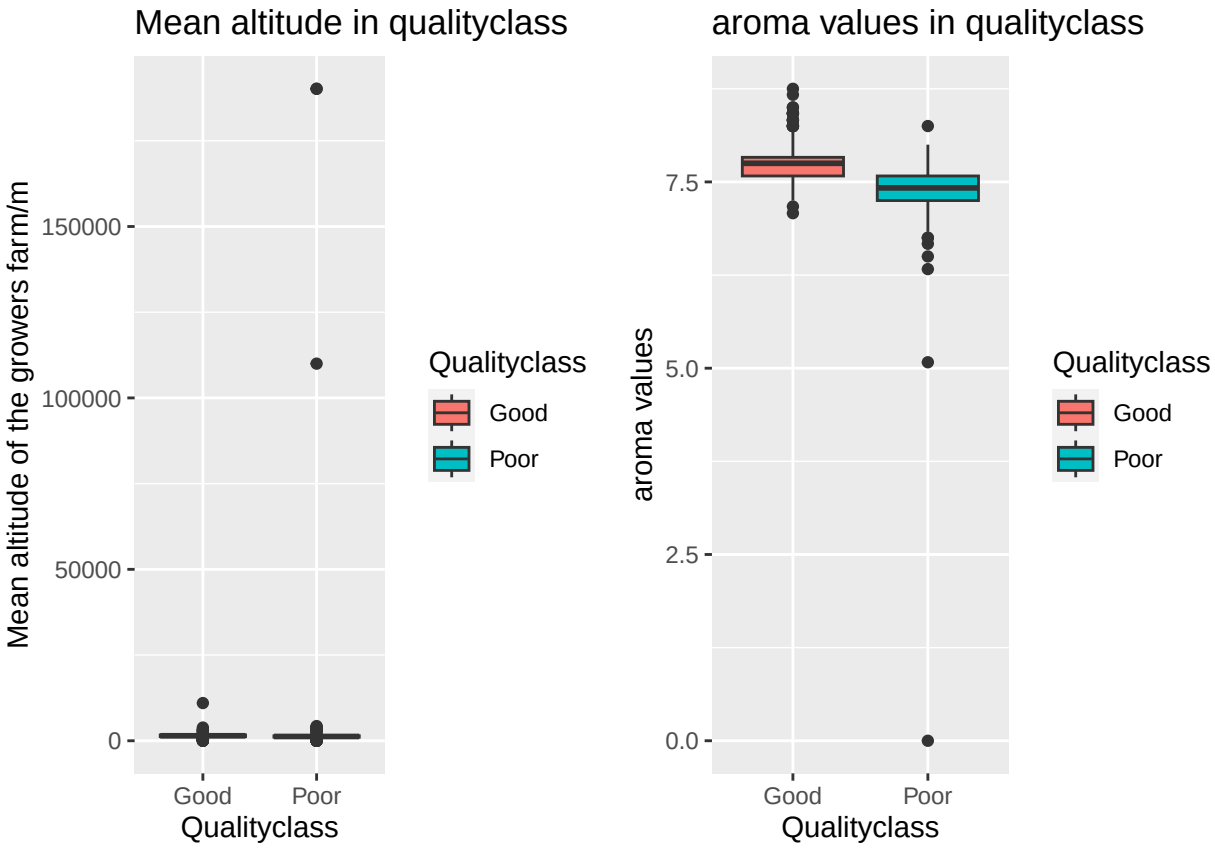

Figure 1: Boxplots of mean altitude of the growers farm/left, aroma values/right in different qualityclass

data cleaning

The boxplot about mean altitude shows that there are some outliers. Four of them are more than 10000 metres, obviously they are wrong datas, so we remove them.From the boxplot about Aroma, we find there is a wrong value which equals zero and remove it from the dataset.

Suitable numerical summaries and visualizations

```
#summary of numerical explanatory variables
newdata_summary<- newdataset%>%
  dplyr::select(aroma,flavor,acidity,category_two_defects,altitude_mean_meters)
my_skim <- skim_with(numeric = sfl(hist = NULL),
                     base = sfl(n = length))
my_skim(newdata_summary) %>%
  transmute(Variable=skim_variable, n = n, Mean=numeric.mean, SD=numeric.sd,
            Min=numeric.p0, Median=numeric.p50,  Max=numeric.p100,
            IQR = numeric.p75-numeric.p50) %>%
  kable(caption = '\\label{tab:summaries1} Summary statistics on the different numerical explanatory va
  kable_styling(font_size = 10, latex_options = "HOLD_position")
```

Table 1: Summary statistics on the different numerical explanatory variables of coffee.

Variable	n	Mean	SD	Min	Median	Max	IQR
aroma	930	7.58	0.31	5.08	7.58	8.75	0.17
flavor	930	7.53	0.32	6.17	7.58	8.67	0.17
acidity	930	7.53	0.31	5.25	7.50	8.58	0.25
category_two_defects	930	3.64	5.35	0.00	2.00	47.00	2.00
altitude_mean_meters	930	1325.65	484.31	1.00	1310.64	4287.00	289.36

Table1 shows that the mean values of Aroma grade, Flavor grade and Acidity grade are both approximately 7.5. There are large differences of category two defects between different coffee beans, as some have no defective product, but some have 47 in the batch of coffee beans tested. Similarly, the difference in mean altitude is distinct.

```
#summary of categorical explanatory variables
#country of origin
data_country<- newdataset %>%
  group_by(country_of_origin) %>%
  summarise(n=n())
data_country
```

```
# A tibble: 33 x 2
   country_of_origin     n
   <chr>             <int>
 1 Brazil               91
 2 Burundi               2
 3 China                14
 4 Colombia            127
```

```
 5 Costa Rica           36
 6 Cote d?Ivoire         1
 7 Ecuador               2
 8 El Salvador          18
 9 Ethiopia             23
10 Guatemala           127
# ... with 23 more rows
```

```
newdataset %>%
  tabyl(country_of_origin, Qualityclass) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() %>%
  kable(caption = '\\label{tab1:origin} Summary statistics on country of origin.') %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 2: Summary statistics on country of origin.

country_of_origin	Good	Poor
Brazil	51.6% (47)	48.4% (44)
Burundi	50.0% (1)	50.0% (1)
China	64.3% (9)	35.7% (5)
Colombia	82.7% (105)	17.3% (22)
Costa Rica	55.6% (20)	44.4% (16)
Cote d?Ivoire	0.0% (0)	100.0% (1)
Ecuador	50.0% (1)	50.0% (1)
El Salvador	72.2% (13)	27.8% (5)
Ethiopia	100.0% (23)	0.0% (0)
Guatemala	52.0% (66)	48.0% (61)
Haiti	20.0% (1)	80.0% (4)
Honduras	26.1% (12)	73.9% (34)
India	50.0% (5)	50.0% (5)
Indonesia	57.1% (8)	42.9% (6)
Kenya	90.0% (18)	10.0% (2)
Laos	0.0% (0)	100.0% (2)
Malawi	9.1% (1)	90.9% (10)
Mauritius	0.0% (0)	100.0% (1)
Mexico	26.0% (52)	74.0% (148)
Myanmar	0.0% (0)	100.0% (6)
Nicaragua	23.1% (3)	76.9% (10)
Panama	75.0% (3)	25.0% (1)
Peru	0.0% (0)	100.0% (1)
Philippines	40.0% (2)	60.0% (3)
Taiwan	40.4% (23)	59.6% (34)
Tanzania, United Republic Of	48.3% (14)	51.7% (15)
Thailand	57.1% (8)	42.9% (6)
Uganda	76.7% (23)	23.3% (7)
United States	66.7% (6)	33.3% (3)
United States (Hawaii)	100.0% (1)	0.0% (0)
United States (Puerto Rico)	33.3% (1)	66.7% (2)
Vietnam	57.1% (4)	42.9% (3)
Zambia	0.0% (0)	100.0% (1)

The summary table shows that there are total 33 countries in the dataset, and 200 observations are from Mexico, which is the most, but some countries have only one observation. We also note that there are 6 countries like Laos only have poor qualityclass of coffee, the qualityclass of Ethiopia and United States(Hawaii) are all good. There also have 3 countries' qualityclass is half and half.

```
#harvested
data_harvested<- newdataset %>%
  group_by(harvested) %>%
  summarise(n=n())
data_harvested
```

```
# A tibble: 9 x 2
  harvested     n
      <dbl> <int>
1      2010    26
2      2011    30
3      2012   255
4      2013   134
5      2014   194
6      2015   118
7      2016   103
8      2017    52
9      2018    18
```

```
newdataset %>%
  tabyl(harvested, Qualityclass) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() %>%
  kable(caption = '\\label{tab1:harvested} Summary statistics on
harvested.') %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 3: Summary statistics on harvested.

harvested	Good	Poor
2010	76.9% (20)	23.1% (6)
2011	73.3% (22)	26.7% (8)
2012	39.2% (100)	60.8% (155)
2013	53.7% (72)	46.3% (62)
2014	50.0% (97)	50.0% (97)
2015	54.2% (64)	45.8% (54)
2016	55.3% (57)	44.7% (46)
2017	48.1% (25)	51.9% (27)
2018	72.2% (13)	27.8% (5)

The summary table shows that the information is collected from 2010 to 2018, and 255 observations is from 2012 which is the most. We also note that in 2010 the propotion of good qualityclass is highest, which is 76.9%. The lowest is 39.2% in 2012.

Country of origin

```
ggplot(newdataset, aes(x=Qualityclass, y=..prop.., group=country_of_origin, fill=country_of_origin))+
  geom_bar(position = "dodge", stat="count")+
  labs(y="Proportion")
```

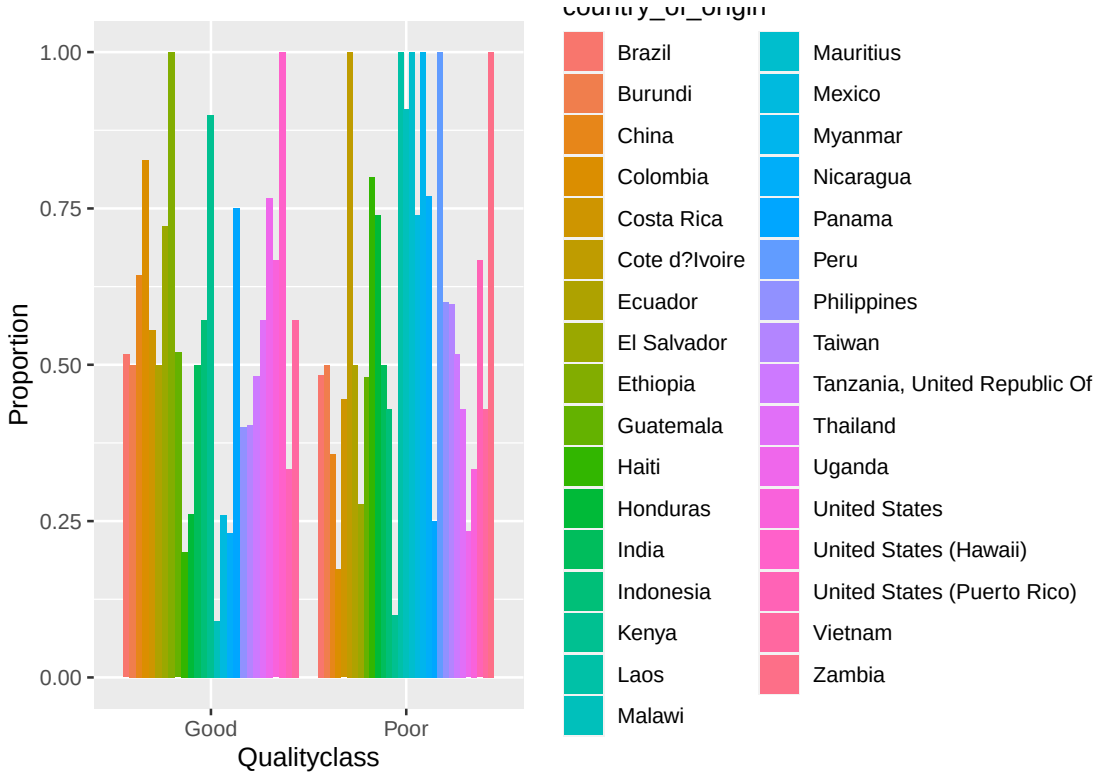

Figure 2: Propotion of Qualityclass by countries of origin.

The plot shows the propotion of Good qualityclass vs poor qualityclass between different country of origin. we can see some countries are all good quality, some are all poor quality.So we can fit a logistic regression model to determine whether the qualityclass of coffee can be predicted from their country of origin.

Aroma, Flavour and Acidity

```
#boxplot of Aroma grade
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = aroma, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Aroma",
       title = "Aroma in different qualityclass")
```

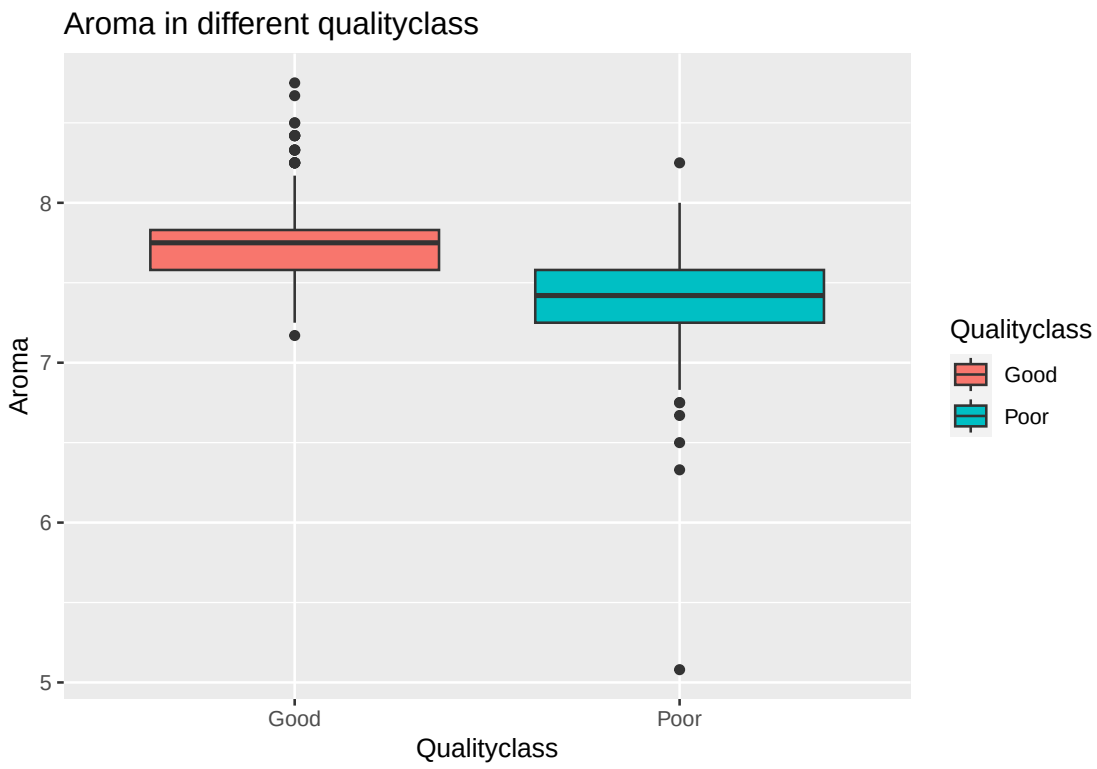

Figure 3: Boxplot of aroma in different qualityclass

```
#boxplot of Flavour grade
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = flavor, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Flavor",
       title = "Flavor in different qualityclass")
```

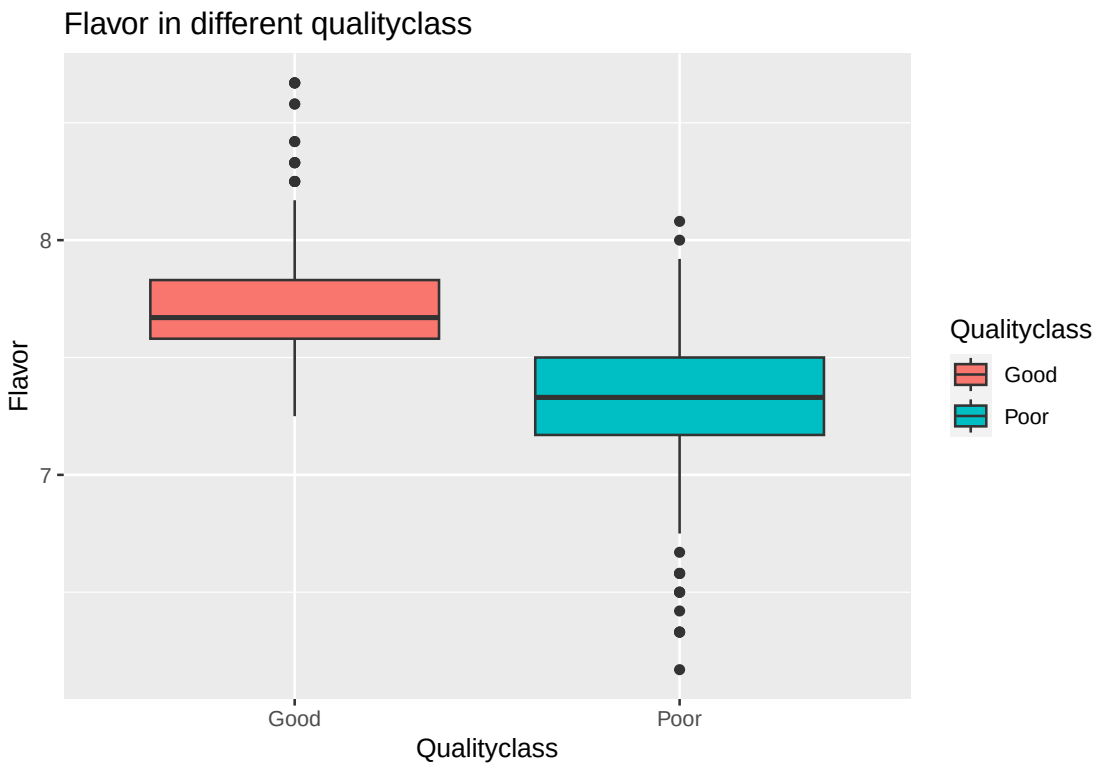

Figure 4: Boxplot of aroma in different qualityclass

```
#boxplot of Acidity grade
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = acidity, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Acidity",
       title = "Acidity in different qualityclass")
```

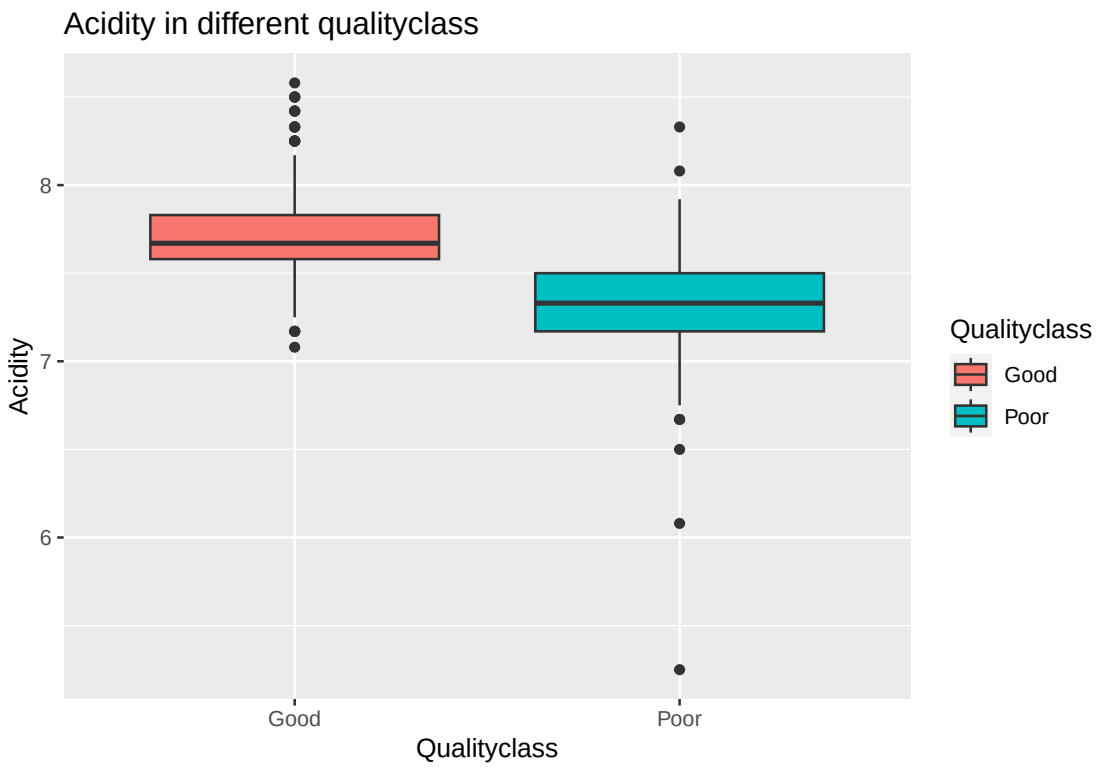

Figure 5: Boxplot of aroma in different qualityclass

The features of these three quality scores are similar. The boxplots show that coffee with good quality have higher grade in Aroma, Flavour and Acidity than poor. So we can fit a logistic regression model to see whether Aroma, Flavour and Acidity are significant predictors of the odds of qualityclass of coffee beans.

Count of defects

```
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = category_two_defects, fill = Quali
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Count of category 2 type defects",
       title = "Count of category 2 type defects in different qualityclass")
```

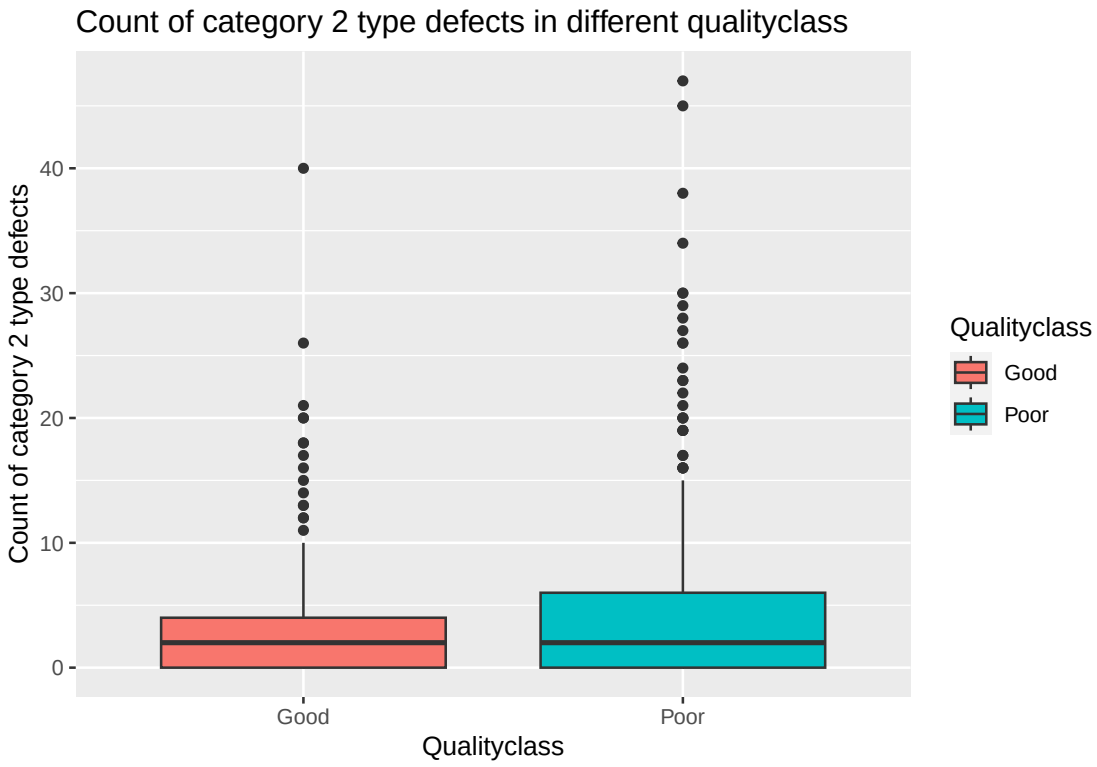

Figure 6: Boxplot of Count of category 2 type defects in different qualityclass

The boxplot shows that the poor quality coffee beans have more defective products than good quality ones, and there are more outliers in poor quality coffee beans.So we can fit a logistic regression model to see whether count of category 2 type defects is a significant predictor of the odds of qualityclass of coffee beans.

Mean altitude

```
ggplot(data = newdataset, mapping = aes(x = factor(Qualityclass), y = altitude_mean_meters, fill = Quali
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Mean altitude of the growers farm/m",
       title = "Mean altitude of the growers farm in different qualityclass")
```

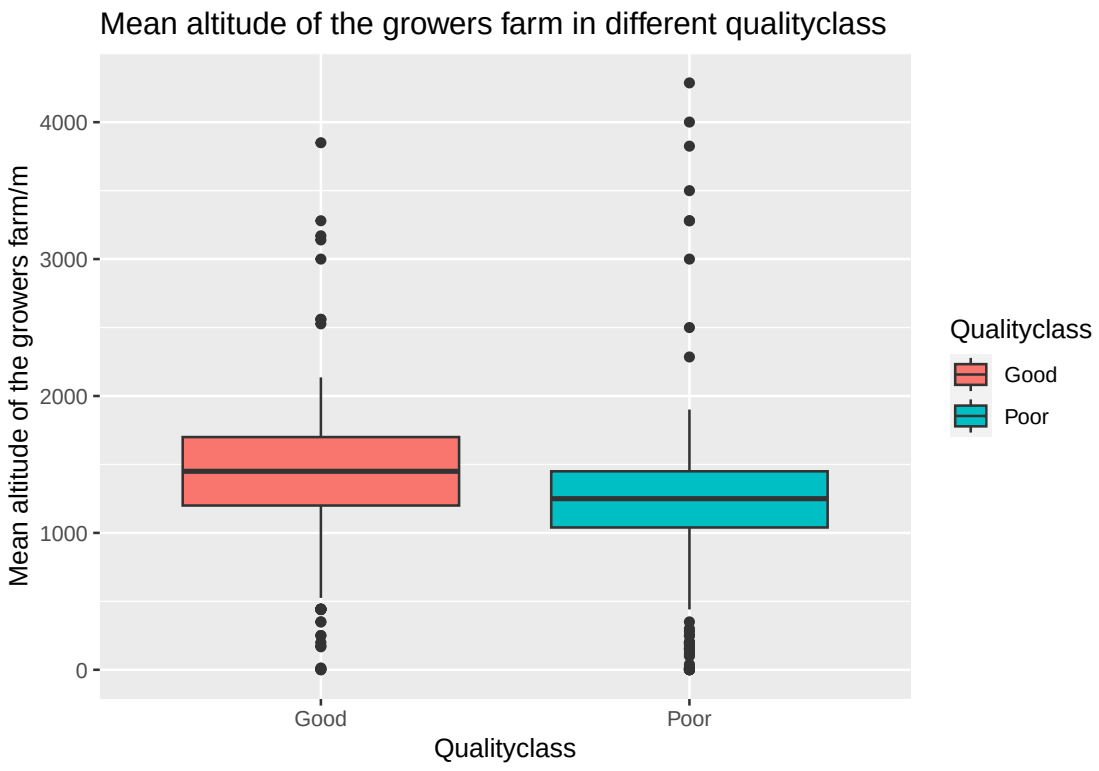

Figure 7: Boxplot of mean altitude of the growers farm in different qualityclass

The boxplot shows that the mean altitude of good quality coffee beans are higher than poor quality ones. we can notice that the poor quality have more outliers. So we can fit a logistic regression model to see whether mean altitude of the growers is a significant predictor of the odds of qualityclass of coffee beans.

harvested

```
ggplot(newdataset, aes(x=Qualityclass, y=..prop.., group=harvested, fill=harvested))+
  geom_bar(position = "dodge", stat="count")+
  labs(y="Proportion")
```

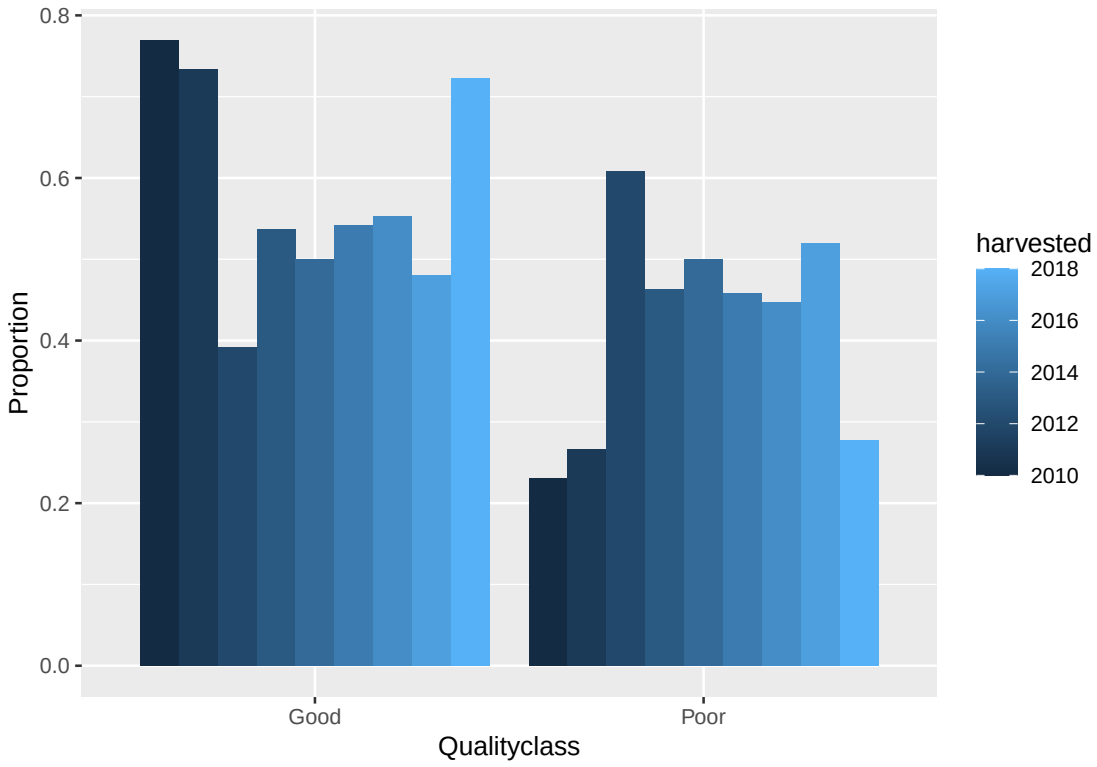

Figure 8: Propotion of Qualityclass by Harvested.

```
prop<-newdataset %>%
  tabyl(harvested, Qualityclass) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns()
prop$Good<-str_sub(string=prop$Good, start=1, end=5)
prop$Good<-as.factor(prop$Good)
ggplot(prop, aes(x=harvested, y=Good, group=1))+
  geom_line()+
  labs(y="Proportion")
```

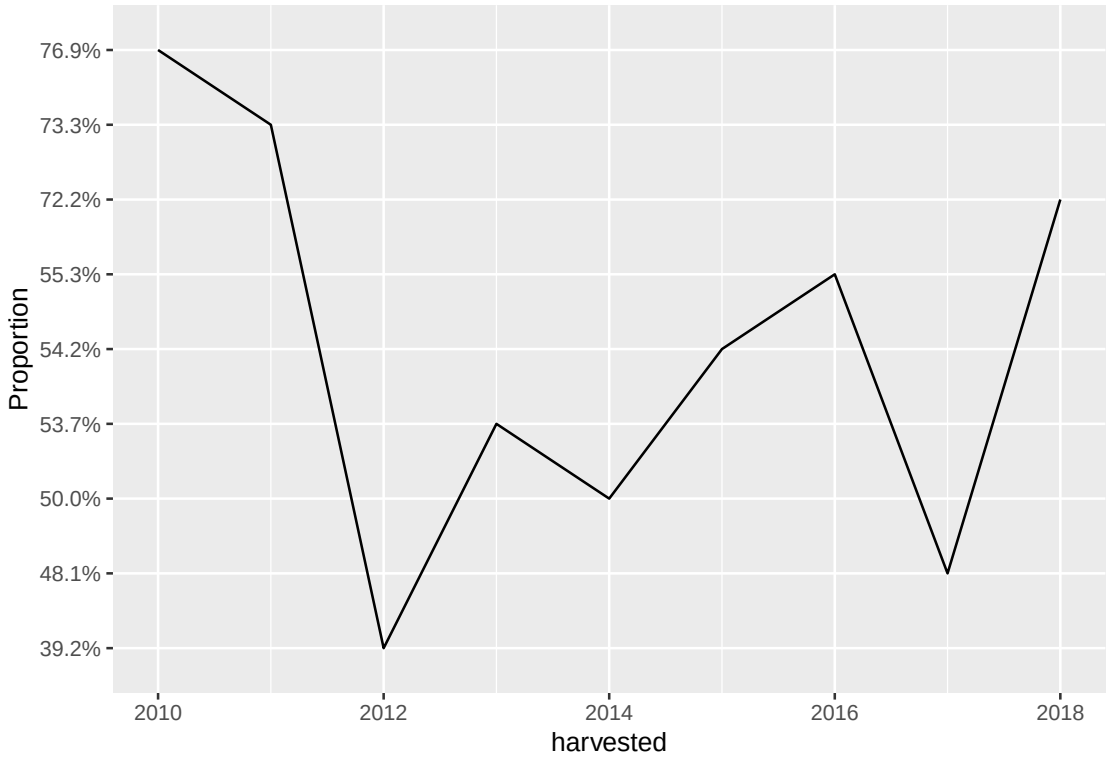

Figure 9: Propotion of Qualityclass by Harvested.

The line plot shows the propotion of Good qualityclass is highest in 2010,which is about 75%. We can fit a logistic regression model to determine whether the qualityclass of coffee can be predicted from harvested years.

Calculating Correlation

In order to prepare for subsequent improvement and selection of variables during modelling, we firstly calculated the correlation between every two numerical variables.

```
newdataset[,2:6]%>%
  cor()%>%
  kable(caption='\\label{tab:correlation} correlation between 5 numerical variables')%>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 4: correlation between 5 numerical variables

	aroma	flavor	acidity	category_two_defects	altitude_mean_meters
aroma	1.0000000	0.7253135	0.5907547	-0.1934092	0.1632542
flavor	0.7253135	1.0000000	0.7438336	-0.2477485	0.1476604
acidity	0.5907547	0.7438336	1.0000000	-0.1851076	0.1778057
category_two_defects	-0.1934092	-0.2477485	-0.1851076	1.0000000	-0.0025717
altitude_mean_meters	0.1632542	0.1476604	0.1778057	-0.0025717	1.0000000

Table 4 shows the correlation between every two variables including aroma, flavor, acidity, category_two_defects and altitude_mean_meters. We can see that the correlation between aroma& flavor

(0.725) and the correlation between flavor&acidity (0.744) are both more than 0.7, which means these pairs have strong positive correlation. There is also a moderate correlation between aroma&acidity (0.591), while the correlation between other pairs are relatively weak.

Processing Non-numerical Data

For non-numerical data, including country_of_origin, Qualityclass and harvested(year), we set the country_of_origin and harvested(year) as factors. While as a qualitative variable, we converted Qualityclass into dummy variables, 'poor' to '0' and 'good' to '1'.

```
names(newdataset)
newdataset$country_of_origin<- as.factor(newdataset$country_of_origin)
newdataset$Qualityclass<- ifelse(newdataset$Qualityclass=='Poor',0,1)
newdataset$harvested <- as.factor(newdataset$harvested)
```

Formal Data Analysis

We used GLM to fit a logistic regression model with Qualityclass as the binary response variable, and country_of_origin, aroma, flavor, acidity, category_two_defects, altitude_mean_meaters and harvested as the explanatory variables. A summary of the model and a graph showing the points estimate for the log-odds with their corresponding 95% confidence interval are obtained as results.

Basic GLM

```
mod.cafe <- glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects +
                  altitude_mean_meters + harvested, data = newdataset, family = binomial(link = "logit"))
print(summary(mod.cafe)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested,
    family = binomial(link = "logit"), data = newdataset)
```

```
tidy(mod.cafe)
```

```
# A tibble: 46 x 5
   term                          estimate std.error statistic  p.value
   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                    -145.       11.6   -12.4     2.15e-35
 2 country_of_originBurundi          1.91      4.93    0.387   6.98e- 1
 3 country_of_originChina            0.500     1.08    0.462   6.44e- 1
 4 country_of_originColombia         1.82      0.564   3.22    1.29e- 3
 5 country_of_originCosta Rica       0.290     0.763   0.380   7.04e- 1
 6 country_of_originCote d?Ivoire  -12.1    6523.     -0.00186 9.99e- 1
 7 country_of_originEcuador         -1.43      1.50   -0.954   3.40e- 1
 8 country_of_originEl Salvador      0.541     0.958   0.565   5.72e- 1
 9 country_of_originEthiopia        13.3     898.      0.0148  9.88e- 1
10 country_of_originGuatemala       -0.583     0.546  -1.07    2.85e- 1
# ... with 36 more rows
```

```
AIC(mod.cafe)
```

```
[1] 543.691
```

```
#AIC = 543
```

##Plot of distribution

```
plot_model(mod.cafe, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.5)
```

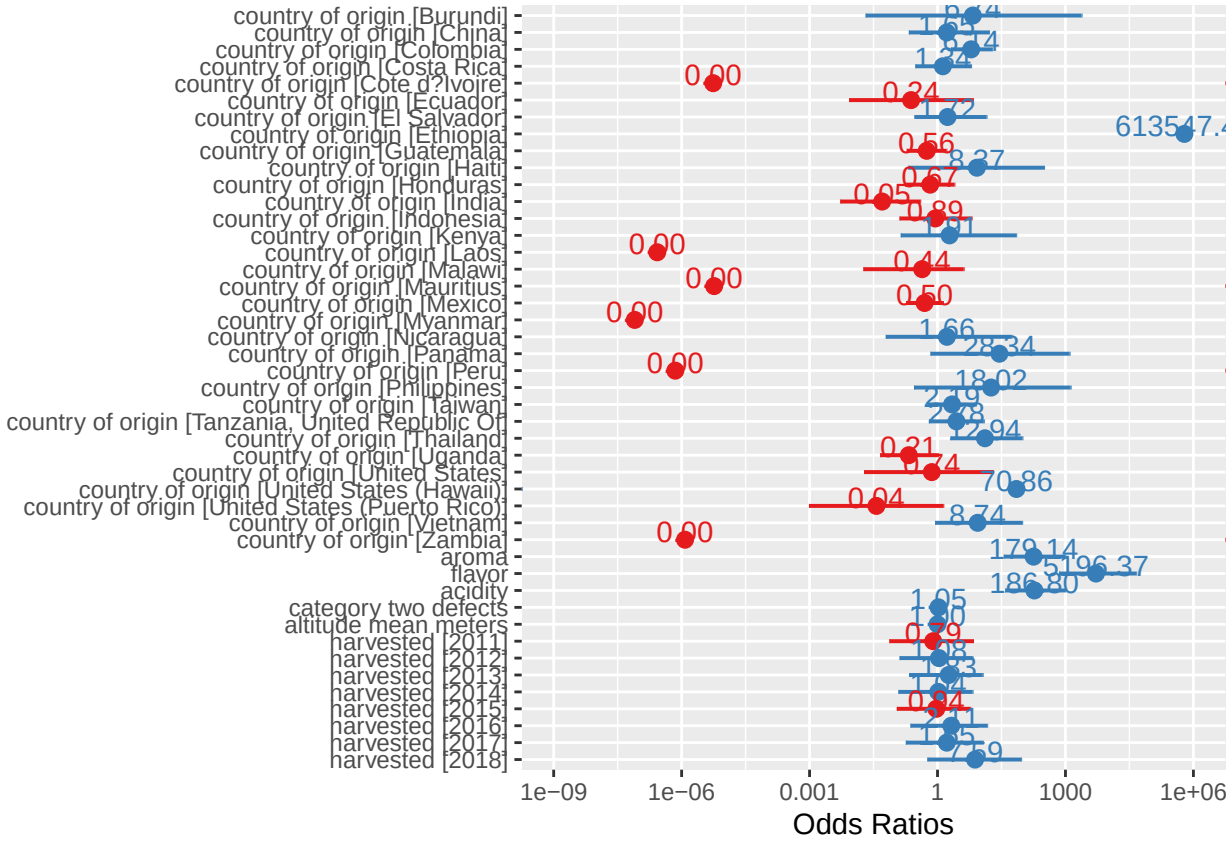

Figure 10: Odds of various factors influencing the quality of coffee(basic GLM model)

In the results we can see that aroma, flavor and acidity has coefficients of 5.19, 8.56, 5,23 separately, indicating comparatively strong positive influence on cafe quality, whilst category_two_defects and altitude_mean_meters do not appear to have much impact. For country_of_origin and harvested, different countries and vintages have different degrees of influence on the quality of coffee. Varies from Thailand(2.56) to India(-2.99). In addition, only the harvested of 2018 shows a little positive impact on cafe quality(2.03), while other variables do not appear to be strongly influential.

GLM Stepwise

In the previous basic GLM we fitted a model with AIC of 543, wondering whether there is better regression to fit the data after selecting only the influencial variables, we then decided to use stepwise regression to improve our model.

```
# Print the selected model
print(summary(model.step)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters, family = binomial(link = "logit"),
    data = newdataset)
```

```
tidy(model.step)
```

```
# A tibble: 38 x 5
   term                           estimate std.error statistic  p.value
   <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -139.        11.0   -12.7     5.64e-37
 2 country_of_originBurundi           1.78       5.23    0.341   7.33e- 1
 3 country_of_originChina            -0.0540     1.01   -0.0535  9.57e- 1
 4 country_of_originColombia          1.59       0.525   3.03    2.45e- 3
 5 country_of_originCosta Rica        0.0645     0.714   0.0903  9.28e- 1
 6 country_of_originCote d?Ivoire   -12.1     6523.     -0.00185 9.99e- 1
 7 country_of_originEcuador          -1.38       1.48   -0.938   3.48e- 1
 8 country_of_originEl Salvador       0.468      0.942   0.497   6.19e- 1
 9 country_of_originEthiopia         13.5      886.       0.0152  9.88e- 1
10 country_of_originGuatemala        -0.776      0.497  -1.56    1.18e- 1
# ... with 28 more rows
```

```
AIC(model.step)
```

```
[1] 537.6971
```

```
#AIC = 537
plot_model(model.step, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.50)
```

Using stepwise regression, we fitted a model with AIC of 537, which is relatively smaller than 543 in our first basic model, hence we can say that stepwise regression helped us to improve our model.

Adding Interaction Terms

Considering the possible interactions between the variables, based on the previously calculated correlations,we added some interaction terms(aroma:flavor,flavor:acidity,aroma:acidity)in order to improve our model.We summarized the results and graphically showed the log-odds and their corresponding 95% confidence intervals.

```
mod.cafe <- glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + al
                  family = binomial(link = "logit"))
print(summary(mod.cafe)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
    acidity + category_two_defects + altitude_mean_meters + harvested +
    aroma:flavor + flavor:acidity + aroma:acidity, family = binomial(link = "logit"),
    data = newdataset)
```

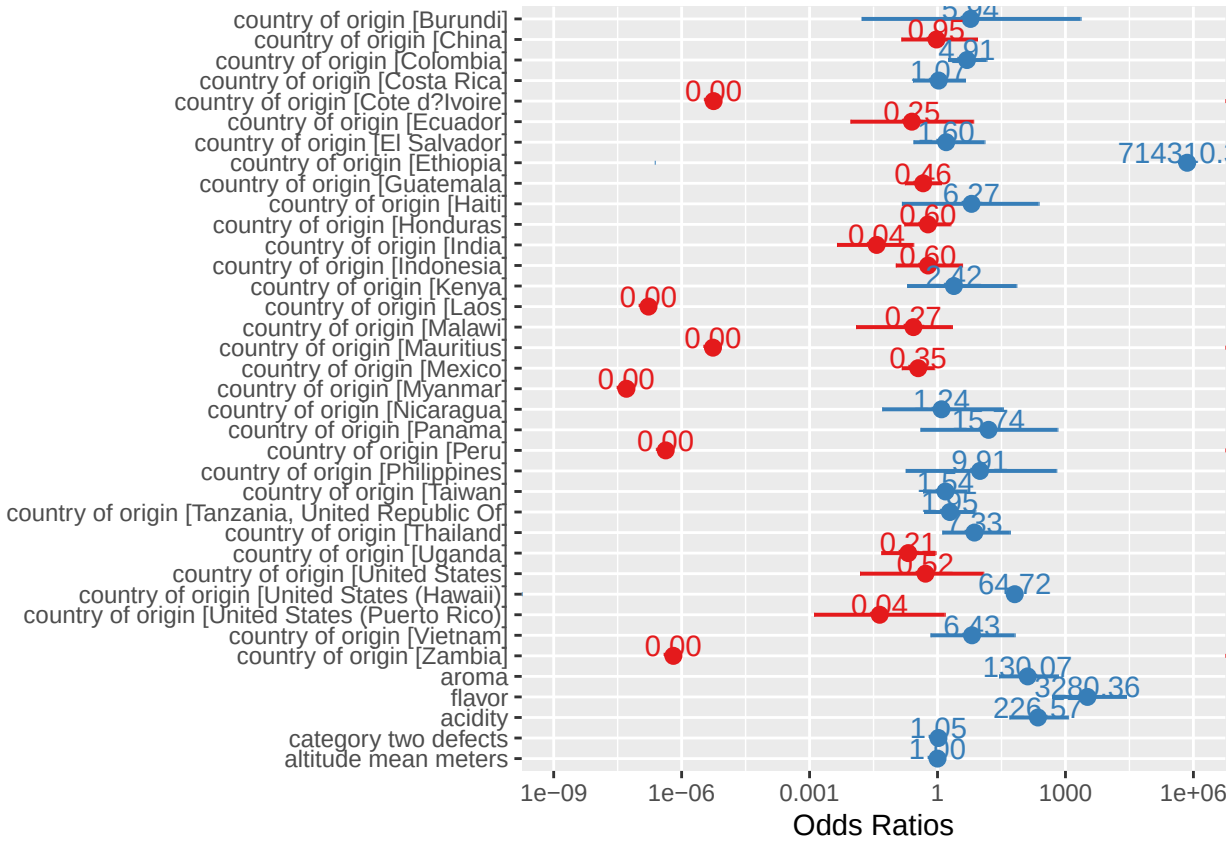

Figure 11: Odds of various factors influencing the quality of coffee(basic GLM model)

```
tidy(mod.cafe)
```

```
# A tibble: 49 x 5
   term                          estimate std.error statistic  p.value
   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                   -1185.       312.     -3.80    0.000144
 2 country_of_originBurundi          3.30       9.72    0.339   0.734
 3 country_of_originChina            0.498      1.08    0.462   0.644
 4 country_of_originColombia         1.78       0.582   3.06    0.00224
 5 country_of_originCosta Rica       0.0985     0.774   0.127   0.899
 6 country_of_originCote d?Ivoire  -10.4     6523.     -0.00159 0.999
 7 country_of_originEcuador         -1.51       1.51   -1.00    0.317
 8 country_of_originEl Salvador      0.543      0.968   0.561   0.575
 9 country_of_originEthiopia        14.6      617.      0.0236  0.981
10 country_of_originGuatemala       -0.591      0.557  -1.06    0.288
# ... with 39 more rows
```

```
AIC(mod.cafe)
```

```
[1] 539.8582
```

```
#AIC = 539
```

```
plot_model(mod.cafe, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.5)
```

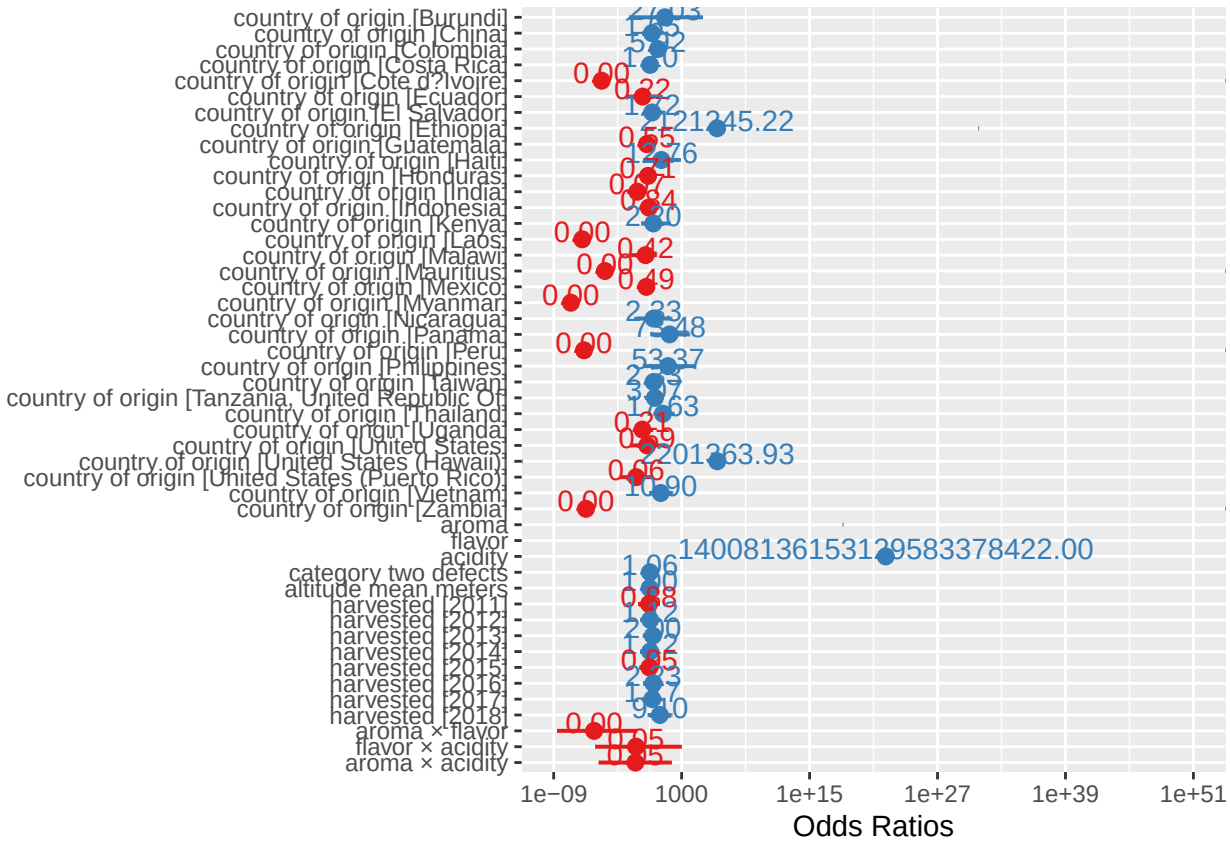

Figure 12: Odds of various factors influencing the quality of coffee(model with interaction terms)

In the results we can see the coefficients of aroma, flavor and acidity themselves are significantly positive, while the coefficients of all our possible interaction terms are negative,which shows that these three variables may moderate each other.After adding interaction terms, we can find that the AIC of the model decreases compared to the basic model, thus we can assume that the addition of the interaction terms improved our model.

GLM Stepwise After Adding Interaction Terms

In order to further improve our model, we fitted the GLM model with interaction terms using the method of stepwise regression with AIC as the criterion.

```
model.step <- stepAIC(glm(Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_de
```

```
print(summary(model.step)$call)
```

```
glm(formula = Qualityclass ~ country_of_origin + aroma + flavor +
```

```
    acidity + category_two_defects + altitude_mean_meters + aroma:flavor,
    family = binomial(link = "logit"), data = newdataset)
```

```
tidy(model.step)
```

```
# A tibble: 39 x 5
   term                           estimate std.error statistic  p.value
   <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -856.       237.     -3.61    0.000305
 2 country_of_originBurundi           2.91       8.89    0.328   0.743
 3 country_of_originChina            -0.0960     1.00   -0.0957  0.924
 4 country_of_originColombia          1.55       0.533   2.92    0.00354
 5 country_of_originCosta Rica       -0.0852     0.711  -0.120   0.905
 6 country_of_originCote d?Ivoire   -11.3     6523.     -0.00173 0.999
 7 country_of_originEcuador          -1.43       1.48   -0.966   0.334
 8 country_of_originEl Salvador       0.575      0.943   0.610   0.542
 9 country_of_originEthiopia         13.5      892.      0.0151  0.988
10 country_of_originGuatemala        -0.766      0.503  -1.52    0.128
# ... with 29 more rows
```

```
AIC(model.step)
```

```
[1] 532.1229
```

```
#AIC = 532
plot_model(model.step, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.50)
```

We can see from the results that the AIC decreased to the lowest among these four models we fitted. AS AIC balances simplicity and accuracy when evaluating models, we can say that after adding an interaction term and doing the stepwise regression, our fourth model is the best model. Also, the last model has the lowest BIC=720, while the other three are 766,721,776 separately, which further demonstrates the superiority of our model.

```
 [1] "Brazil"                       "Burundi"
 [3] "China"                        "Colombia"
 [5] "Costa Rica"                   "Cote d?Ivoire"
 [7] "Ecuador"                      "El Salvador"
 [9] "Ethiopia"                     "Guatemala"
[11] "Haiti"                        "Honduras"
[13] "India"                        "Indonesia"
[15] "Kenya"                        "Laos"
[17] "Malawi"                       "Mauritius"
[19] "Mexico"                       "Myanmar"
[21] "Nicaragua"                    "Panama"
[23] "Peru"                         "Philippines"
[25] "Taiwan"                       "Tanzania, United Republic Of"
[27] "Thailand"                     "Uganda"
[29] "United States"                "United States (Hawaii)"
[31] "United States (Puerto Rico)"  "Vietnam"
[33] "Zambia"
```

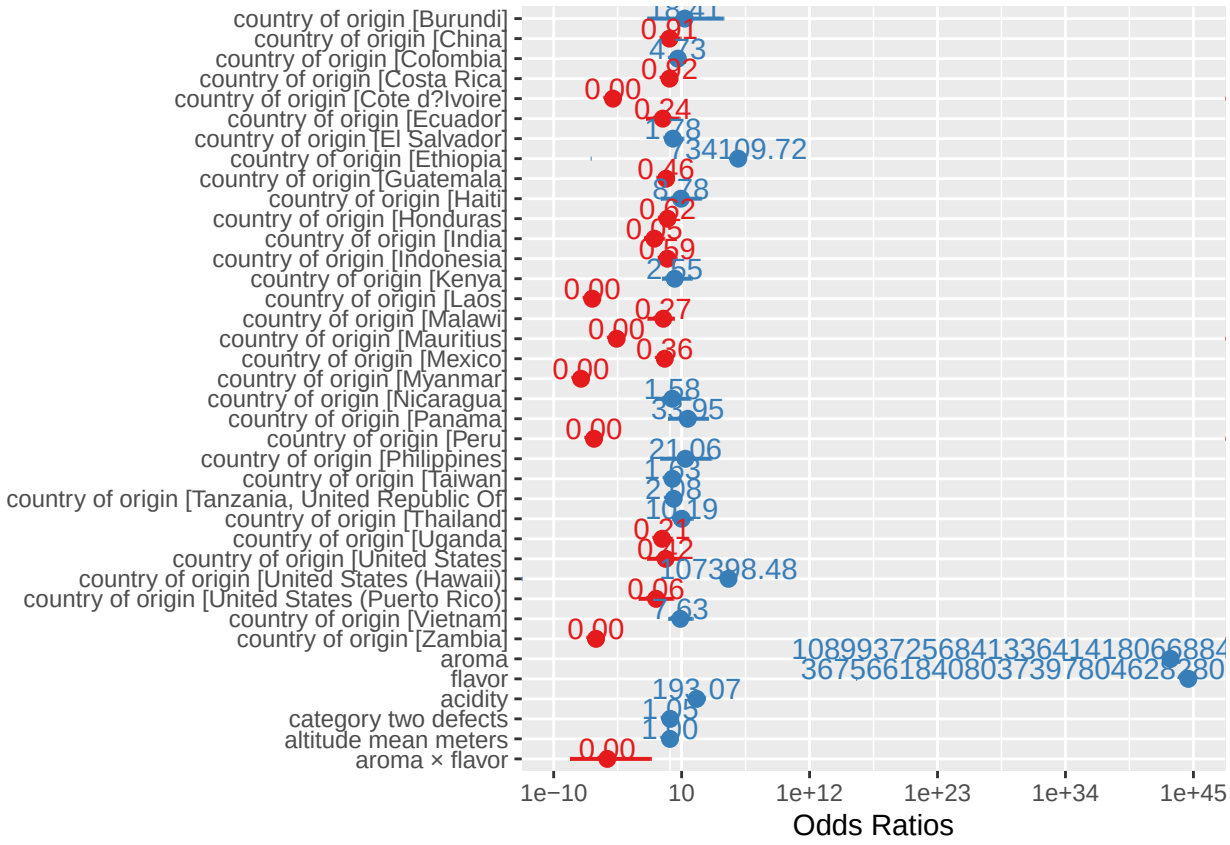

Figure 13: Odds of various factors influencing the quality of coffee(stepwise regression with interaction terms)

checking assumptions

Residuals Plots for each variables

We see that there is an even spread of the residuals above and below the zero line for each variables,although there are a very few outlier points, overrall their spread on the graphs are acceptable, hence our assumption that the residuals have mean zero appears valid.

Density Histogram

In the graph we can see that the residuals are normally distributed with the mean 0, therefore the assumption is valid.

The remaining assumptions hold naturally at the time of our modelling, thus our model appears valid.

Conclusion

After data cleaning and processing of non-numerical data, we fitted the data to a regression model to observe the effect of each variable in the dataset on coffee quality, and we continued to improve the model by stepwise

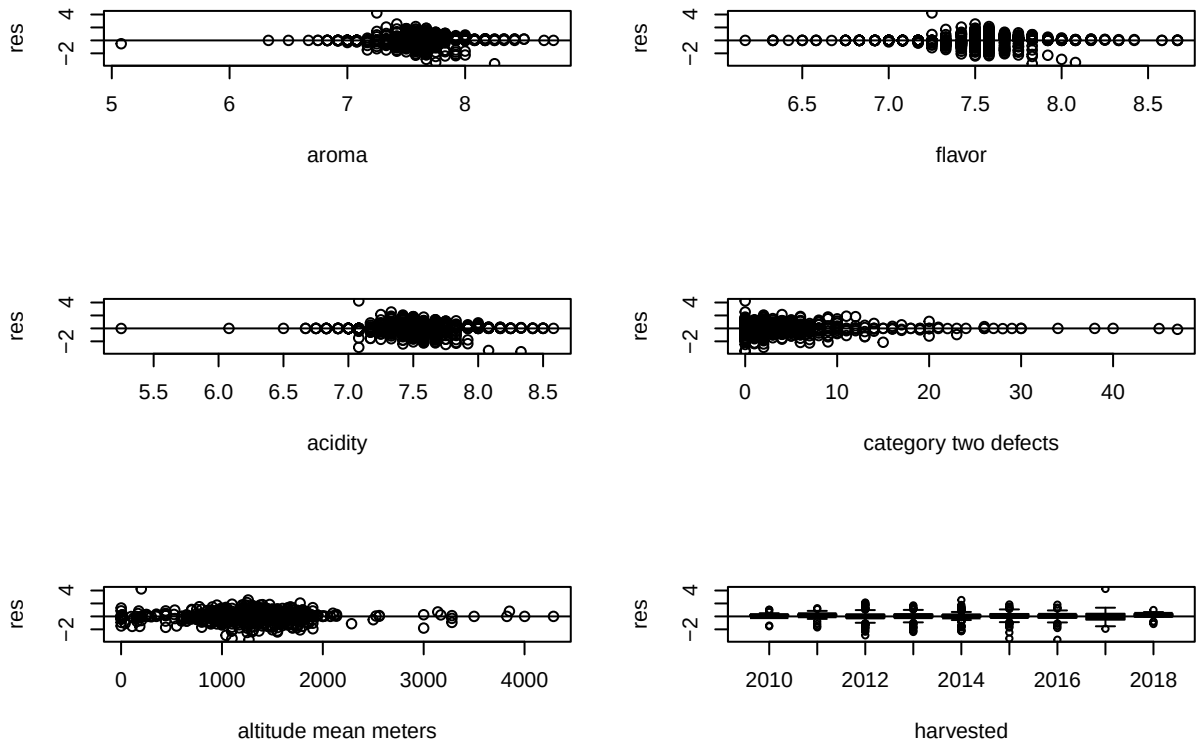

Figure 14: residuals against each variables

regression and adding possible interaction terms, resulting in the model with the smallest AIC value and therefore the most profile accurate model4. Looking at the summaries and graphs of model4, we can pick out the factors that have the greatest impact: aroma and flavor are very positively influencing on the quality of coffee, with coefficients of 99.1 and 102.62. The influence of origin varies very much. When the p-value is less than a certain level of significance (0.05), which means H0 is rejected, we can see that Colombia(1.56) and Thailand(2.32) have the highest coefficient of all countries, while Uganda(-1.56), India(-3.09), Mexico(-1.03) have less coefficient, making relatively negative effects. However, many remaining origins do not seem to have obvious impact on the quality of the coffee.

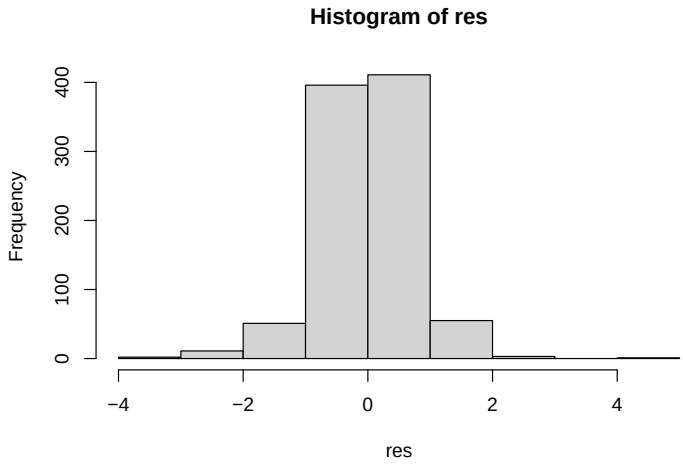

Figure 15: density histogram of residuals